Nonlin. Processes Geophys., 21, 1145–1157, 2014
www.nonlin-processes-geophys.net/21/1145/2014/
doi:10.5194/npg-21-1145-2014

Nonlinear Processes in Geophysics
Open Access

Non-parametric Bayesian mixture of sparse regressions with application towards feature selection for statistical downscaling

D. Das[1,2], J. Dy[3], J. Ross[3], Z. Obradovic[2], and A. R. Ganguly[1]

[1]Sustainability and Data Sciences Lab, Northeastern University, Boston, MA, USA
[2]Center for Data Analytics and Biomedical Informatics, Temple University, Philadelphia, PA, USA
[3]Department of Electrical and Computer Engineering, Northeastern University, Boston, MA, USA

Correspondence to: A. R. Ganguly (a.ganguly@neu.edu)

Received: 27 February 2014 – Published in Nonlin. Processes Geophys. Discuss.: 11 April 2014
Revised: 21 August 2014 – Accepted: 23 October 2014 – Published: 1 December 2014

Abstract. Climate projections simulated by Global Climate Models (GCMs) are often used for assessing the impacts of climate change. However, the relatively coarse resolutions of GCM outputs often preclude their application to accurately assessing the effects of climate change on finer regional-scale phenomena. Downscaling of climate variables from coarser to finer regional scales using statistical methods is often performed for regional climate projections. Statistical downscaling (SD) is based on the understanding that the regional climate is influenced by two factors – the large-scale climatic state and the regional or local features. A transfer function approach of SD involves learning a regression model that relates these features (predictors) to a climatic variable of interest (predictand) based on the past observations. However, often a single regression model is not sufficient to describe complex dynamic relationships between the predictors and predictand. We focus on the covariate selection part of the transfer function approach and propose a nonparametric Bayesian mixture of sparse regression models based on Dirichlet process (DP) for simultaneous clustering and discovery of covariates within the clusters while automatically finding the number of clusters. Sparse linear models are parsimonious and hence more generalizable than non-sparse alternatives, and lend themselves to domain relevant interpretation. Applications to synthetic data demonstrate the value of the new approach and preliminary results related to feature selection for statistical downscaling show that our method can lead to new insights.

1 Introduction

Climate change is one of most challenging problems facing humankind. Its impacts are expected to influence policy decisions on critical infrastructures, management of natural resources, humanitarian aid, and emergency preparedness along with numerous regional-scale human economic and social activities. Therefore, it is imperative to accurately assess the impacts of climate change at regional scale in order to inform stakeholders for appropriate decision making related to mitigation policies. Global climate models (GCMs) are the most credible tools at present for future climate projection that accounts for the effects of greenhouse gas emissions under different socio-economic scenarios. Although GCMs perform reasonably well in projecting climate variables at a larger spatial scale ($> 10^4$ km^2), they perform poorly for regional-scale climate projections. Such poor performance of the GCMs, coupled with the importance of regional climate projections for impact studies, has led to development of limited area models (LAMs) or regional climate models (RCMs), where finer spatial grids over a limited spatial area are embedded within a coarser GCM grid. This method is also known as dynamic downscaling. However, these models are complex, computationally expensive and require rerunning for each new region. Moreover, regional models inherit the basic gaps in understanding of climate physics that limit the performance of GCMs. A couple of recently published studies (Kumar et al., 2014; Knutti and Sedláček, 2013) rigorously compared the projections of the latest generation of climate models (CMIP5) with the previous generation (CMIP3) but found no significant improvement in the

majority of statistical performance metrics even with higher spatial resolutions and addition of new physical processes in the computational model. Uncertainties in sub-grid-scale cloud-microphysics and ocean eddy processes and poor understanding of the effect of carbon cycle and other biogeochemical processes on climate systems still limit the ability of the physics-based climate models to reliably project future climate (Bader et al., 2008), especially at regional scale.

A complementary approach for regional projection is statistical downscaling that uses statistical models to learn empirical statistical relationships between large-scale GCM features (predictors) and regional-scale climate variable(s) (predictands) to be projected. The statistical approaches of downscaling can be categorized into three broad classes – weather typing, weather generators, and the transfer function approaches (Wilby et al., 2004). Weather typing approaches have originally been developed for weather forecasting and generally involve classifying days into similar clusters or weather states based on their synoptic similarity. Typically, weather patterns are clustered based on their similarity with nearest neighbors while the statistical models they use vary in their definition of similarity measures. On the other hand, weather generators replicate the statistical properties of the daily predictand variable by using a stochastic model, such as Markov processes (Greene et al., 2011), that uses wet–dry and dry–wet transition probabilities as input for training while conditioning its parameters on large-scale predictors.

In this paper, however, we are interested in transfer function based regression models that learn a linear or nonlinear mapping between large scale predictors and regional scale predictand variables. Regression models are conceptually the simplest of the three classes since they provide a direct mapping between the predictor and predictand values. However, the success of the regression models depends on the accurate choice of predictors. Sparse regressions based on constrained L1-norm (Tibshirani, 1994) of the coefficients became popular due to their ability to simultaneously select covariates and fit parsimonious linear models that are more generalizable and easily interpretable. Although sparse regression models have been applied widely in many disciplines, their application to climate, and especially to statistical downscaling, has remained very limited. In a recent paper (Ebtehaj et al., 2012), sparse regularization has been shown to be effective for downscaling rainfall fields for weather forecasting, whereas sparse variable selection has been used for statistical downscaling of climate variables (Phatak et al., 2011) in a separate paper. To our knowledge, there is no other published work on use of sparse regularization for statistical downscaling.

However, large complex climate data sets often exhibit dynamic behavior (Kannan and Ghosh, 2010) which may not be modeled well by a single regression model. Here we propose a nonparametric model for mixture of sparse regressions that can accommodate multiple sparse linear relationships inherent in the data set. Nonparametric models are more flexible than the finite mixture models (Bishop and Svenskn, 2002) since they assume no prior knowledge about the number of distinct components in the data. We used a Dirichlet process mixture (DPM) (Antoniak, 1974) with stick-breaking construction (Ishwaran and James, 2001) to accommodate an unknown number of sparse regression models in the data. DPM start by assuming infinite components in the data but ends up discovering a finite number of components supported by the data. We used the Bayesian version of sparse regression (Park and Casella, 2008) to smoothly integrate the sparse regression model with the DPM, which is a nonparametric Bayesian approach where each component is represented by a set of distribution parameters specific to the corresponding component.

Although the number of different components may not be known, prior knowledge often exists about whether a pair of observations belong to the same component. For example, it is reasonable to assume that two observations close in time from the same location may exhibit similar behavior. We allow soft "must link" constraints between pairs of data-points that encourage the pair to belong to the same mixture component. Such constraints are incorporated in our Bayesian model with the help of a Markov random field (MRF) prior over the cluster indicator variables (Ross and Dy, 2013; Basu et al., 2006).

Variational Bayesian (VB) inference has been shown to be much faster than stochastic alternatives for nonparametric Bayesian models (Blei and Jordan, 2006). The major contribution of this paper is to develop a fully Bayesian formulation for nonparametric mixture of a sparse regression model and designing an efficient variational inference algorithm to obtain posterior distributions over the regression coefficients of potentially multiple regression components as well as the component membership probabilities of each data-point.

We have extensively demonstrated the performance of our algorithm on synthetic data. We have also applied our method to the feature selection problem for statistical downscaling of annual average rainfall over two regions on the west coast of the USA. Preliminary results from the application of our algorithm to select features for regression based statistical downscaling show that our method may lead to improved prediction and discovery of new insights.

2 Background

In this section, we provide brief descriptions of the methods in the context they were used to build our model.

2.1 Bayesian sparse regression

Let us assume that we are given a data set $D = \{\boldsymbol{x}_n,\ y_n : n = 1, \ldots N\}$ that has been generated from a linear model identified by sparse coefficients vector $\boldsymbol{\beta}$. In a non-Bayesian setting, sparsity is enforced by a constraint on the L1-norm of

the coefficients which is given by

$$\boldsymbol{y}_n = \boldsymbol{\beta}^\top \boldsymbol{x}_n + \epsilon, \quad \text{subject to } ||\boldsymbol{\beta}||_1^1 \leq t, \tag{1}$$

where $\epsilon \sim \mathcal{N}(0,\, \tau^{-1})$.

However, in a Bayesian setting, the sparsity can be imposed by a Laplace prior (also known as double exponential distribution) on $\boldsymbol{\beta}$ which is given by (Park and Casella, 2008):

$$p(\boldsymbol{\beta}|\boldsymbol{\gamma}, \tau) = \prod_{j=1}^{D} \frac{\sqrt{\gamma_j \tau}}{2} \exp\left(-\sqrt{\gamma_j \tau}|\beta_j|\right). \tag{2}$$

However, due to the analytical intractability of the Laplace prior, it is often represented in the following scale-mixture (of Gaussians) form using an additional random variable $\boldsymbol{\alpha}$.

$$\begin{aligned} p(\boldsymbol{\beta}|\tau,\, \boldsymbol{\gamma}) &= \prod_{j=1}^{D} \frac{\sqrt{\gamma_j \tau}}{2} \exp\left(-\sqrt{\gamma_j \tau}|\beta_j|\right) \\ &= \prod_{j=1}^{D} \int \mathcal{N}\left(\beta_j; 0, \tau^{-1}\alpha_j^{-1}\right) \\ &\quad \mathrm{InvGa}\left(\alpha_j; 1, \frac{\gamma_j}{2}\right) \mathrm{d}\alpha_j \end{aligned}$$

For a fully hierarchical Bayesian setting, Gamma prior is imposed on parameter τ as well as on individual penalty parameters γ_j. So the joint distribution over all the parameters can be given by

$$\begin{aligned} p(\boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = \mathrm{Ga}(\tau; c_0, d_0) \prod_{j=1}^{D} \Big\{ &\mathcal{N}\left(\beta_j; , 0, \tau^{-1}\alpha_j^{-1}\right) \\ &\times \mathrm{InvGa}\left(\alpha_j; 1, \frac{\gamma_j}{2}\right) \mathrm{Ga}\left(\gamma_j; a_0, b_0\right) \Big\}. \end{aligned} \tag{3}$$

2.2 Markov random fields

An MRF is represented by an undirected graphical model in which the nodes represent variables or groups of variables and the edges indicate dependence relationships. An important property of MRFs is that a collection of variables is conditionally independent of all others in the field given the variables in their Markov blanket. The Hammersley–Clifford theorem states that the distribution, $p(\boldsymbol{Z})$, over the variables in an MRF factorizes according to

$$p(\boldsymbol{Z}) = \frac{1}{\mathcal{Z}} \exp\left(-\sum_{c \in C} H_{\mathrm{c}}(\boldsymbol{z}_{\mathrm{c}})\right), \tag{4}$$

where $\mathcal{Z}$ is a normalization constant called the *partition function*, C is the set of all cliques in the MRF, $\boldsymbol{z}_{\mathrm{c}}$ is the set of variables in clique c, and H_{c} is the *energy function* over clique c (Geman and Geman, 1984). A clique is a set of nodes in a graph that are fully connected. The smallest clique in a graph is an edge. The energy function captures the desired configuration of local variables. *Partition function* $\mathcal{Z}$ normalizes the probability measure and it is computed by summing the exponentiated energy functions of all possible configurations.

2.3 Dirichlet process mixture (DPM)

The Dirichlet process (DP) was first introduced in statistics literature as a measure on measures (Ferguson, 1973). It is parameterized by a base measure, G_0, and a positive scaling parameter λ:

$$G|\{G_0, \lambda\} \sim \mathrm{DP}(G_0, \lambda). \tag{5}$$

The notion of a DPM arises if we treat the kth draw from G as a parameter of the distribution over some observation (Antoniak, 1974) representing a particular mixture component. DPMs can be interpreted as mixture models with an infinite number of mixture components in the sense that data exhibit a finite number of components but previously unseen components represented by new data can still be accommodated. More recently, a variational inference algorithm for DPMs was introduced (Blei and Jordan, 2006) using the stick-breaking construction (Sethuraman, 1994) which uses two infinite collections of random variables $V_k \sim \mathrm{Beta}(1,\, \lambda)$ and $\boldsymbol{\eta}_k^* \sim G_0$ to construct G as

$$\theta_k = V_k \prod_{j=1}^{k-1} (1 - V_j) \tag{6}$$

$$G(\boldsymbol{\eta}) \sim \sum_{k=1}^{\infty} \theta_k \delta\left(\boldsymbol{\eta}, \boldsymbol{\eta}_k^*\right). \tag{7}$$

For a mixture of sparse regression models, if the parameters for each components are given by $\boldsymbol{\eta}_k$, the subsequent data generation process for such a mixture model can be described in the following steps using a stick-breaking construction:

1. Draw $v_k \sim \mathrm{Beta}(1, \lambda)\ k = \{1,\ 2,\ \ldots \infty\}$
2. Draw $\boldsymbol{\eta}_k \sim G_0$, $k = \{1,\ 2,\ \ldots \infty\}$
3. Generate $\theta_k = v_k \prod_{m=1}^{k-1} (1 - v_m)$.
4. For each data-point n:
   a. Draw $\boldsymbol{z}_n \sim \mathrm{Mult}(\boldsymbol{\theta})$
   b. Draw $y_n \sim \mathcal{N}(y_n;\ \boldsymbol{x}_n,\ \boldsymbol{\eta}_{z_n})$.

We can truncate the construction process at $k = K$ by enforcing $v_{K-1} = 0$ which forces all θ_k for $k > K$ to be zero (see step 3). The resulting construction is called a truncated DP (TDP), which can be shown to approximate the true DP quite well given K is large relative to the number of the datapoints (Ishwaran and James, 2001).

3 Methodology

Now, let us assume that we are given a data set $D = \{\boldsymbol{x}_n,\ y_n : n = 1,\ \ldots N\}$ which has been generated from a mixture of

K different sparse models identified by sparse coefficients $\boldsymbol{\beta}^{(1)}$, $\boldsymbol{\beta}^{(2)}$, …, $\boldsymbol{\beta}^{(K)}$. Let us also assume that the number of components K is unknown. We use a Bayesian formulation of the sparse regression model for each component $\boldsymbol{\beta}^{(k)}$, with $k = 1, 2, \ldots K$. Let us first state the Bayesian version of the kth sparse model. The linear regression model of the kth component can be represented by the following Gaussian distribution.

$$p\left(y_n|\boldsymbol{x}_n, \boldsymbol{\beta}^{(k)}\right) \sim \mathcal{N}\left(y_n; \boldsymbol{\beta}^{(k)\top}\boldsymbol{x}_n, \tau_k^{-1}\right) \tag{8}$$

3.1 Mixture of sparse regressions

We introduce K-dimensional latent indicator variables $\{z_n : n = 1, \ldots N\}$ to represent the component membership of each data-point $\{\boldsymbol{x}_n,\ y_n\}$. If the data-point belongs to the kth component, then z_{nk} will be 1 and all other elements of $\boldsymbol{z}_n$ will be 0. We further denote $\boldsymbol{Z} = [z_1\, z_2 \ldots z_n]$. We can now rewrite Eq. (8) in terms of $\boldsymbol{z}_n$ as

$$p\left(y_n|\boldsymbol{x}_n, \left\{\boldsymbol{\beta}^{(k)}\right\}\right) \sim \prod_{k=1}^{K}\left\{\mathcal{N}\left(y_n; \boldsymbol{\beta}^{(k)\top}\boldsymbol{x}_n, \tau_k^{-1}\right)\right\}^{z_{nk}}. \tag{9}$$

For this mixture of sparse regressions model, each component has a separate parameter set $\{\boldsymbol{\beta}^{(k)},\ \tau_k\}$. Moreover, after adding the parameters related to the scale-mixture representation of the Laplace prior on $\boldsymbol{\beta}^{(k)}$ (refer to Sect. 2.1), the set of parameters is finally given by $\boldsymbol{\eta}_k = \{\boldsymbol{\beta}^{(k)},\ \tau_k,\ \boldsymbol{\alpha}_k,\ \boldsymbol{\gamma}_k\}$. The prior distribution G_0 from which these parameters can be drawn jointly is given in Eq. (3). We can now use the stick-breaking construction described in Sect. 2.3 to formulate our mixture model. The overall generative process is then:

$$\begin{aligned}
&p\left(\boldsymbol{y}, \mathbf{Z}, \boldsymbol{v}, \left\{\boldsymbol{\beta}^{(k)}\right\}, \boldsymbol{\tau}, \left\{\boldsymbol{\alpha}^{(k)}\right\}, \left\{\boldsymbol{\gamma}^{(k)}\right\}, \lambda|\mathbf{X}\right)\\
&= p\left(\boldsymbol{y}|\mathbf{X}, \left\{\boldsymbol{\beta}^{(k)}\right\}, \boldsymbol{\tau}\right) p(\mathbf{Z}|\boldsymbol{v}) p(\boldsymbol{v}|\lambda) p\left(\lambda|m_0\right)\\
&\times p\left(\left\{\boldsymbol{\beta}^{(k)}\right\}|\boldsymbol{\tau}, \left\{\boldsymbol{\alpha}^{(k)}\right\}\right) p\left(\left\{\boldsymbol{\alpha}^{(k)}\right\}|\left\{\boldsymbol{\gamma}^{(k)}\right\}\right)\\
&\times p\left(\left\{\boldsymbol{\gamma}^{(k)}\right\}|a_0, b_0\right) p\left(\boldsymbol{\tau}|c_0, d_0\right).
\end{aligned} \tag{10}$$

The graphical model that represents the dependence relationships between all the parameters involved in this current mixture model is shown in Fig. 1. The shaded circles denote observed variables; the unshaded circles denote unobserved variables. We have used a Gamma prior on λ having a hyper-parameter m_0. We have omitted the hyper-parameters a_0, b_0, c_0, d_0, and m_0 from the list of conditioning variables in the left side to avoid clutter. The individual distributions in Eq. (10) are given below.

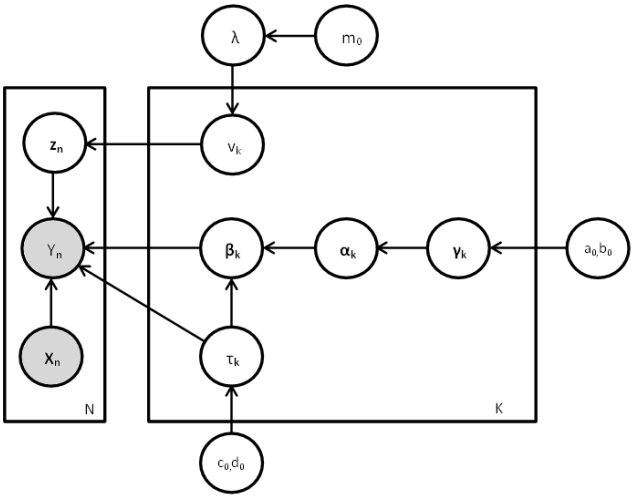

Figure 1. Graphical representation of the complete Bayesian hierarchical model.

$$\boldsymbol{y}|\mathbf{X}, \left\{\boldsymbol{\beta}^{(k)}\right\}, \boldsymbol{\tau} \sim \prod_{n=1}^{N}\prod_{k=1}^{K}\left\{\mathcal{N}\left(y_n; \boldsymbol{x}_n^{\top}\boldsymbol{\beta}^{(k)}, \tau_k^{-1}\right)\right\}^{z_{nk}} \tag{11a}$$

$$\mathbf{Z}|\boldsymbol{v} \sim \prod_{n=1}^{N}\prod_{k=1}^{K}\left\{v_k\prod_{j=1}^{k-1}\left(1-v_j\right)\right\}^{z_{nk}} \tag{11b}$$

$$\boldsymbol{v}|\lambda \sim \prod_{k=1}^{K}\mathrm{Beta}\left(v_k; 1, \lambda\right) \tag{11c}$$

$$\lambda \sim \mathrm{Ga}\left(\lambda; m_0, 1\right) \tag{11d}$$

$$\left\{\boldsymbol{\beta}^{(k)}\right\}|\tau_k, \left\{\boldsymbol{\alpha}^{(k)}\right\} \sim \prod_{k=1}^{K}\prod_{j=1}^{D}\mathcal{N}\left(\beta_j^{(k)}; 0, \left(\tau_k\alpha_j^{(k)}\right)^{-1}\right) \tag{11e}$$

$$\boldsymbol{\tau} \sim \prod_{k=1}^{K}\mathrm{Ga}\left(\tau_k; c_0, d_0\right) \tag{11f}$$

$$\begin{aligned}
\left(\left\{\boldsymbol{\alpha}^{(k)}\right\}, \left\{\boldsymbol{\gamma}^{(k)}\right\}\right) \sim &\prod_{k=1}^{K}\prod_{j=1}^{D}\mathrm{InvGa}\left(\alpha_j^{(k)}; 1, \frac{\gamma_j^{(k)}}{2}\right)\\
&\times \mathrm{Ga}\left(\gamma_j^{(k)}; a_0, b_0\right)
\end{aligned} \tag{11g}$$

3.2 Accommodating "must link" constraints

Prior knowledge about must link constraints between pairs of data-points can be enforced via an MRF prior on the indicator variables $\boldsymbol{z}_n$, where each data-point is considered a node and each constraint between a pair of data-points is regarded as an edge between the respective nodes. We denote the collection of edges by $\mathcal{C}$ and the MRF prior is given by Eq. (4). We define the energy function as:

$$\mathrm{H}\left(z_i, z_j\right) = \begin{cases} -1, & z_i^{\top}z_j = 1 \text{ and } (i,j) \text{ is ML} \\ 0, & \text{otherwise} \end{cases}. \tag{12}$$

Here ML means must link. This prior encourages similar values of indicator variables z_i and z_j if they happen to share a "must link" edge. Since the MRF prior is assigned only on the indicator variables $\mathbf{Z}$, it only alters Eq. (11b) and the new prior on $\mathbf{Z}$ is given by

$$\mathbf{Z}|\boldsymbol{v} \sim \frac{1}{\mathcal{Z}} \exp\left(-\sum_{(i,j)\in\mathcal{C}} H\left(z_i, z_j\right)\right) \times \prod_{n=1}^{N}\prod_{k=1}^{K}\left\{v_k \prod_{j=1}^{k-1}\left(1-v_j\right)\right\}^{z_{nk}}. \quad (13)$$

3.3 Variational inference

Let us consider all the unknown parameters in our model as latent variables and denote all the latent variables by $\mathbf{H} = \{\mathbf{Z},\ \boldsymbol{v},\ \{\boldsymbol{\beta}^{(k)}\},\ \boldsymbol{\tau},\ \{\boldsymbol{\alpha}^{(k)}\},\ \{\boldsymbol{\gamma}^{(k)}\},\ \lambda\}$. Moreover, from now on, we will ignore feature variables $\mathbf{X}$ from the list of conditioning variables as they are observed. Using Jensen's inequality, we can find a lower bound of the log-marginal $\ln p(\boldsymbol{y})$ which is given as

$$\ln p(\boldsymbol{y}) > \int q(\mathbf{H}) \ln\left\{\frac{p(\boldsymbol{y},\mathbf{H})}{q(\mathbf{H})}\right\} \mathrm{d}\mathbf{H} \quad (14)$$

for any arbitrary distribution $q(\mathbf{H})$. The variational inference is performed by restricting $q(\mathbf{H})$ within a parametric family so that the maximization of the lower bound given in Eq. (14) is tractable. We consider only those $q(\mathbf{H})$ that factorize over some disjoint groups of the component random variables of $\mathbf{H}$ in the following way:

$$q(\mathbf{H}) = \prod_{j=1}^{L} q_j\left(\boldsymbol{h}_j\right). \quad (15)$$

We can now maximize the lower bound given in Eq. (14) with respect to each component $q_j(\boldsymbol{h}_j)$ in Eq. (15) and obtain the parametric form of $q_j(\boldsymbol{h}_j)$ given by

$$q_j^*\left(\boldsymbol{h}_j\right) = \frac{\exp\left(\mathrm{E}_{i\neq j}\left[\ln p\left(\boldsymbol{y},\mathbf{H}\right)\right]\right)}{\int \exp\left(\mathrm{E}_{i\neq j}\left[\ln p(\boldsymbol{y},\mathbf{H})\right]\right)\mathrm{d}\boldsymbol{h}_j}, \quad (16)$$

where the expectation is taken with respect to all the other factors $\{q_i\}$ for $i \neq j$. It can be shown that the $q(\mathbf{H})$ obtained this way is the closest approximation of the actual posterior $p(\mathbf{H}|y)$ in terms of KL-divergence out of all possible alternatives of the form given by Eq. (15). Therefore this is a deterministic but approximate posterior inference method, unlike stochastic inference methods such as MCMC, which samples from the actual posterior. However, variational inference is much faster and approximates the true posterior reasonably well for practical purposes.

Once we apply Eq. (16) to the joint distribution described in Eqs. (10) and (11), we can get the update equations for the approximate posterior distributions for each of the latent variables involved.

1. Distribution of $\boldsymbol{z}$:

$$q_Z(\mathbf{Z}) = \prod_{V\in\mathcal{V}}\left[\frac{1}{\mathcal{Z}_V}\exp\left(-\sum_{\substack{(i,j)\in\mathcal{C}\\ i,j\in V}} H\left(z_i,z_j\right)\right) \prod_{n\in V}\prod_{k=1}^{K}\rho_{nk}^{z_{nk}}\right] \quad (17)$$

with

$$\rho_{nk} = \frac{r_{nk}}{\sum_k r_{nk}} \quad (18)$$

$$\ln r_{nk} = \frac{1}{2}\langle\ln\tau_k\rangle - \frac{1}{2}\ln 2\pi - \frac{\langle\tau_k\rangle}{2} \left(y_n^2 - 2\langle\boldsymbol{\beta}^{(k)}\rangle^\top \boldsymbol{x}_n y_n + \boldsymbol{x}_n^\top \langle\boldsymbol{\beta}^{(k)}\left(\boldsymbol{\beta}^{(k)}\right)^\top\rangle \boldsymbol{x}_n\right) + \langle\ln v_k\rangle + \sum_{j=1}^{k-1}\langle\ln\left(1-v_j\right)\rangle. \quad (19)$$

2. Distribution of $\{\boldsymbol{\beta}^{(k)}\}$:

$$q_\beta\left(\left\{\boldsymbol{\beta}^{(k)}\right\}\right) = \prod_{k=1}^{K}\mathcal{N}\left(\left\{\boldsymbol{\beta}^{(k)}\right\}; \boldsymbol{\mu}_k, \Sigma^{(k)}\right) \quad (20)$$

with

$$\Sigma^{(k)} = \left(\langle\tau_k\rangle\sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{x}_n^\top E[\mathbf{Z}]_{nk} + \langle\tau_k\rangle \mathrm{diag}\left(\langle\boldsymbol{\alpha}^{(k)}\rangle\right)\right)^{-1} \quad (21)$$

$$\boldsymbol{\mu}_k = \Sigma^{(k)}\left(\sum_{n=1}^{N}\boldsymbol{x}_n y_n E[\mathbf{Z}]_{nk}\right)\langle\tau_k\rangle. \quad (22)$$

Here $\mathrm{diag}(\langle\boldsymbol{\alpha}^{(k)}\rangle)$ corresponds to the LASSO (Tibshirani, 1994) shrinkage. The moments are given by[1]

$$\langle\beta^{(k)}\rangle = \boldsymbol{\mu}_k;\ \left\langle\left(\beta_p^{(k)}\right)^2\right\rangle = \Sigma_{pp}^{(k)} + \mu_{kp}^2$$

$$\langle\boldsymbol{\beta}^{(k)}\left(\boldsymbol{\beta}^{(k)}\right)^\top\rangle = \Sigma^{(k)} + \boldsymbol{\mu}_k\boldsymbol{\mu}_k^\top.$$

3. Distribution of $\boldsymbol{\tau}$:

$$q_\tau(\boldsymbol{\tau}) = \prod_{k=1}^{K}\mathrm{Ga}\left(\tau_k; c_k, d_k\right) \quad (23)$$

[1] $\langle f(s)\rangle$ means expected value of $f(s)$ with respect to the distribution of s.

with

$$c_k = c_0 + \frac{1}{2}\left(\sum_{n=1}^{N} E[\mathbf{Z}]_{nk} + p\right) \tag{24}$$

$$d = d_0 + \frac{I}{2} + \frac{J}{2} \tag{25}$$

where

$$I = \sum_{n=1}^{N}\Big(y_n^2 E[\mathbf{Z}]_{nk} - 2E[\mathbf{Z}]_{nk}\boldsymbol{x}_n^\top y_n \langle\boldsymbol{\beta}^{(k)}\rangle + E[\mathbf{Z}]_{nk}\boldsymbol{x}_n^\top \langle\boldsymbol{\beta}^{(k)}\left(\boldsymbol{\beta}^{(k)}\right)^\top\rangle \boldsymbol{x}_n\Big)$$

$$J = \sum_{p=1}^{D}\left\langle\alpha_p^{(k)}\right\rangle\left\langle\left(\beta_p^{(k)}\right)^2\right\rangle.$$

The relevant moments are

$$\langle\tau_k\rangle = c_k/d_k \text{ and } \langle\ln\tau_k\rangle = \psi(c_k) - \ln(d_k).$$

4. Distribution of $\boldsymbol{v}$

$$q_v(\boldsymbol{v}) = \prod_{k=1}^{K}\text{Beta}(v_k;\xi_k,\kappa_k) \tag{26}$$

with

$$\xi_k = 1 + \sum_{n=1}^{N} E[\mathbf{Z}]_{nk} \text{ and } \kappa_k = \langle\lambda\rangle + \sum_{j=k+1}^{K}\sum_{n=1}^{N} E[\mathbf{Z}]_{nj}.$$

Relevant moments are given by $\langle\ln v_k\rangle = \psi(\xi_k) - \psi(\xi_k+\kappa_k)$ and $\langle\ln(1-v_k)\rangle = \psi(\kappa_k) - \psi(\xi_k+\kappa_k)$. 5. Distribution of $\{\boldsymbol{\alpha}^{(k)}\}$:

$$q_\alpha\left(\left\{\boldsymbol{\alpha}^{(k)}\right\}\right) = \prod_{k=1}^{K}\prod_{p=1}^{D}\text{InvGaussian}\left(\alpha_p^{(k)};g_p^k,h_p^k\right) \tag{27}$$

with

$$g_j^k = \sqrt{\frac{\left\langle\gamma_j^{(k)}\right\rangle}{\langle\tau_k\rangle\left\langle\left(\beta_j^{(k)}\right)^2\right\rangle}}$$

$$h_j^k = \left\langle\gamma_j^{(k)}\right\rangle$$

where InvGaussian($\alpha_j^{(k)}$; g_j^k, h_j^k) denotes inverse Gaussian distribution with mean g_j^k and shape parameter h_j^k having the following density function.

$$p_{\text{IG}}\left(\alpha_j^{(k)};g_j^k,h_j^k\right) = \sqrt{\frac{h_j^k}{2\pi\left(\alpha_j^{(k)}\right)^3}} \times \exp\left(-\frac{h_j^k\left(\alpha_j^{(k)}-g_j^k\right)^2}{2\left(g_j^k\right)^2\alpha_j^{(k)}}\right)\left(\alpha_j^{(k)}>0\right)$$

The relevant moments are given by

$$\left\langle\alpha_j^{(k)}\right\rangle = g_j^k \text{ and } \left\langle\left(\alpha_j^{(k)}\right)^{-1}\right\rangle = \left(g_j^k\right)^{-1} + \left(h_j^k\right)^{-1}.$$

6. Distribution of $\{\boldsymbol{\gamma}^{(k)}\}$:

$$q_\gamma(\{\boldsymbol{\gamma}^{(k)}\}) = \prod_{p=1}^{D}\text{Ga}\left(\gamma_j^{(k)};a_j^k,b_j^k\right) \tag{28}$$

with

$$a_j^k = a_0 + 1$$

$$b_j^k = b_0 + \frac{1}{2}\left\langle\left(\alpha_j^{(k)}\right)^{-1}\right\rangle$$

and the relevant moment is $\langle\gamma_j^{(k)}\rangle = a_j^k/b_j^k$. 7. Distribution of λ:

$$q_\lambda(\lambda) = \text{Ga}(\lambda;u,w) \tag{29}$$

where

$$u = m_0 + K; \quad w = -\sum_{k=1}^{K}\langle\ln(1-v_k)\rangle.$$

Relevant moment is $\langle\lambda\rangle = \frac{u}{w}$.

The first part of the variational posterior of $q_Z(\mathbf{Z})$ in Eq. (17) arises from the MRF prior and contributes towards enforcing "must link" constraints. Note that $\mathcal{V}$ in Eq. (17) is a set of sets and V is a component set of connected nodes within $\mathcal{V}$. Basically, $\mathcal{V}$ denotes the set of connected components within the constraint graph described in Sect. 3.2. Therefore the partition function $\mathcal{Z}_V$ needs to be computed only for the connected components, not for the entire graph. Computing $\mathcal{Z}_V$ becomes tractable if the connected components are small (i.e., the constraint set is sparse).

In order to automatically generate a sparse constraints set, we first implemented all the constraints in the form of edges and then used a graph partitioning algorithm (Hespanha, 2004) to partition the constraint graph in such a way that none of the partitions are left with more than a predefined number of nodes. At the time of inference we used a "backtracking" algorithm (Tarjan, 1972) to find the strongly connected components within the graph. To compute the expectation $E[z]$, we first computed the multinomial probabilities ρ_{nk} and then did an MRF update on each connected

component by computing the probabilities of each possible state combination and summing the probability-weighted state matrices. The partition function is computed by summing the exponentiated sum of energy function of each state matrix. Note that isolated nodes (not part of any connected components) will not need their ρ_{nk} updated.

The parameters of each of the distributions has dependency on moments of one or more of the other variables. We therefore find a locally optimum solution via an iterative process that starts with random initial values of the relevant moments and stops when the indicator variables $\mathbf{Z}$ stop changing. Note that once the approximate solution is reached, we can compute the marginal distributions over coefficients $\beta_p^{(k)}$ which is a Gaussian with mean $\mu_p^{(k)}$ and variance $\Sigma_{pp}^{(k)}$ for each k. We can thereby perform a t test to determine whether the corresponding feature has a non-zero coefficient.

3.4 Computational considerations

One computational bottleneck of the proposed VB algorithm is the inversion of the $D \times D$ matrix in Eq. (21). If $D < N$, then faster matrix inversion can be achieved by first applying a Cholesky decomposition and then inverting the resulting upper triangular matrix. However, if $D > N$, we can first apply a fast (approximate) singular value decomposition on $\Sigma^{(k)-1}$ and then use Woodbury matrix inversion identity so that we now have to invert a $N \times N$ matrix instead.

We have truncated the infinite DP at $K = 20$ for most of our experiments. The speed of the algorithm can be further improved by parallelizing the updates for each of K components, which is straightforward as they are updated independent of each other. Another major computational challenge was the MRF updates. Apart from controlling the maximum size of the connected components, we parallelized the MRF updates over each subgraph by making the state generation independent of the previous state.

4 Experiments

We have evaluated our method on both synthetic and climate data sets. Typical values used for the hyper-parameters were $a_0 = b_0 = c_0 = d_0 = 0.01$ and $\lambda = 1$. Selecting these values within a reasonable range does not affect the results significantly. We made sure that the cardinality of the largest connected component in the constraints graph never exceeds 8.

4.1 Synthetic data set

We compared the performance of both constrained and unconstrained versions of our method with the non-parametric mixture of linear regression (NPMLR) model without any regularization. We set up three experiments: (1) to test whether or not our algorithm can learn the correct number of clusters; (2) to evaluate the effect of constraints; and (3) to check the sensitivity of our approach to noise.

For all our experiments involving synthetic data, we used $N = 1000$ data-points and $D = 30$ features. In our first set of experiments we tested our method for $K = 2 \ldots 5$ actual clusters. Each column of the $N \times D$ input matrix $\mathbf{X}$ is generated from a uniform distribution. For each value of K, we partitioned the input matrix $\mathbf{X}$ in K equal parts $\mathbf{X}_1 \ldots \mathbf{X}_K$. Then for each partition $\mathbf{X}_k$ ($k = 1 \ldots K$), we generate sparse coefficients $\boldsymbol{\beta}_k$ by randomly selecting 10 out of 30 components to be non-zero. We assign a value of $5k$ (where k is the index of the cluster, $k = 1, \ldots, K$) to the non-zero components within the kth cluster so that two clusters are distinctly identifiable in case the indices of non-zero components of the clusters are the same. We then generate the output $\boldsymbol{y}_k$ for the kth cluster using the linear regression model of Eq. (1). The fixed noise variance τ_k^{-1} for the first experiment was generated by randomly choosing a number between 0 and 0.1 to introduce diversity. A final data set was obtained by merging $\{\mathbf{X}_k, \boldsymbol{y}_k\}$ for all $k = 1 \ldots K$. The process is repeated 30 times and mean and variance of the evaluation metrics were reported in the form of error bars for each value of K in Fig. 2. For all these experiments, the total number of constraints was kept at 20 per cluster while the size of the largest subgraph was kept below 7.

The second experiment was performed to evaluate the effect of number of "must link" constraints on the performance of the constrained version of the algorithm. Here, the actual number of clusters was fixed at $K = 3$ along with the base noise variance (0.1) and the number of constraints per cluster was varied from 0 to 30 incremented by 5, although the actual number of constraints may be less since we removed some constraints to achieve sparsity in the constraint graph. The result is reported in Fig. 3.

In our third experiment, we evaluated the effect of noise on the performance of our algorithm. Again, we kept the number of clusters fixed at $K = 3$ and the number of constraints fixed at 20 per cluster (for the constrained version). We varied the base noise level in each cluster from 0 to 0.5 and added a randomly generated value between 0 and 0.1 with the base noise level for each cluster to maintain diversity among the clusters. Average and variance of 30 repetitions are reported in Fig. 4.

4.1.1 Evaluation metrics

We measured two aspects of the performance of our algorithm. First,we measured whether it can cluster the data-points correctly. We put a data-point into one of the possible 20 components (since we truncated the infinite DP at $K = 20$ for all experiments) depending on the value of the row $E[\mathbf{Z}]_n$ (a vector) in the $N \times 20$ matrix $E[\mathbf{Z}]$ estimated by the variational inference algorithm. The estimated cluster membership $\hat{c}_n$ (a scalar) is given by $\hat{c}_n = \text{argmax}_k E[\mathbf{Z}]_{nk}$. We retain all the valid components out of 20 possible, which have at least one member initially. Then we run an update algorithm to merge very small clusters with the closest larger

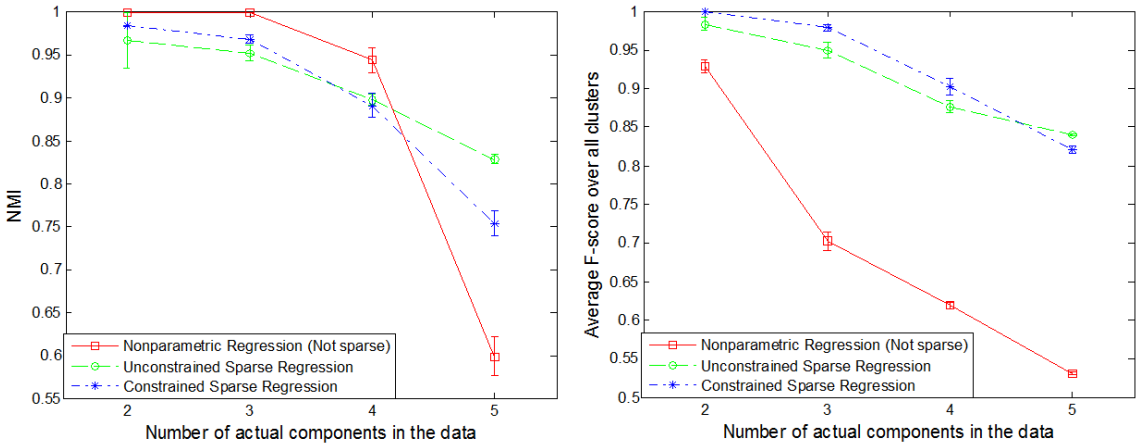

Figure 2. Left panel: ability of nonparametric unregularized and sparse regressions (unconstrained and constrained) to correctly identify clusters in presence of increased number of actual components in the data. Right panel: ability of nonparametric unregularized and sparse regressions (unconstrained and constrained) to correctly retrieve the sparse structure within each cluster.

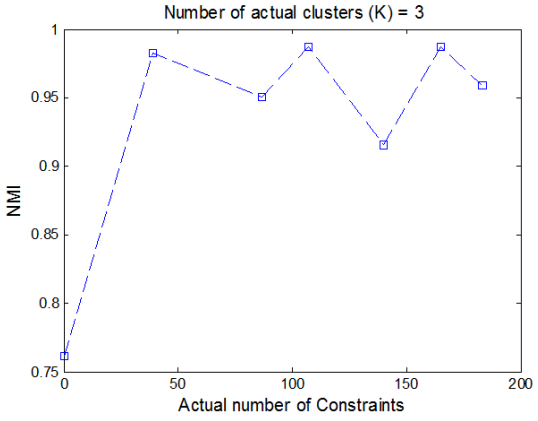

Figure 3. Performance of the constrained version of the algorithm (in terms of NMI (more the better)) with number of "must link" constraints.

ones. Note that the estimated cluster indices (a value between 1 and 20) may not correspond directly to the actual cluster indices (a value between 1 to actual value of K) since the variational inference algorithm is not aware of the actual order of the cluster indices (e.g., actual cluster index 1 may correspond to estimated cluster index 9). So we use a metric called normalized mutual information (NMI) that evaluates the match between estimated cluster memberships $\hat{\boldsymbol{c}}$ and actual ones $\boldsymbol{c}$ without needing direct correspondence. NMI is given by $\mathrm{NMI}(\boldsymbol{c}, \hat{\boldsymbol{c}}) = \frac{H(\boldsymbol{c}) - H(\boldsymbol{c}|\hat{\boldsymbol{c}})}{\sqrt{H(\boldsymbol{c})H(\hat{\boldsymbol{c}})}}$, where $H(\cdot)$ is the entropy. Higher NMI values mean that the clustering results are more similar to ground-truth. The metric reaches its maximum value of one when there is perfect agreement.

A second metric is used to evaluate the quality of the sparse regression model estimated within each discovered cluster. Here we are only interested in finding whether our algorithm picks the non-zero coefficients correctly. We use F score to measure the match between actual and estimated non-zero coefficients within each cluster. F score for the kth component is given by $F_k = \frac{2P_k R_k}{P_k + R_k}$, where P_k is the precision and R_k is the recall of the estimated coefficients for the kth component. We reported the average of F_k values over all components discovered by our algorithm. Unlike the previous metric, here we need to know the direct correspondence between the cluster indices so that we can match the actual and estimated coefficient vectors. We developed an algorithm to find such a correspondence based on bipartite matching.

4.1.2 Discussion of results

We can see the performance of all three algorithms are comparable in terms of identifying the clusters correctly, although the NMI value of NPMLR degrades significantly for $K = 5$. However, as desired, our method outperforms NPMLR in terms of correctly retrieving the sparse structure of regression coefficients within each cluster. There is a general downward trend of performance for all algorithms with increasing number of actual components in the data. This is an inherent problem with the DPM models as it tends to attach each new data-point to the largest current component, thereby favoring models with fewer components. Also, as the number of actual components grows, the probability of two components being similar increases.

The increased flexibility of non-parametric methods comes at a cost of hitting local optima being more likely and finding solutions that are not interpretable. Adding more constraints may decrease this probability but at the same time restricts the variational method from finding solutions leading to a larger lower bound, especially in the presence of more components in the data. Therefore increasing the number of constraints may result in more interpretable solutions, but not improved accuracy. It is also encouraging to see that

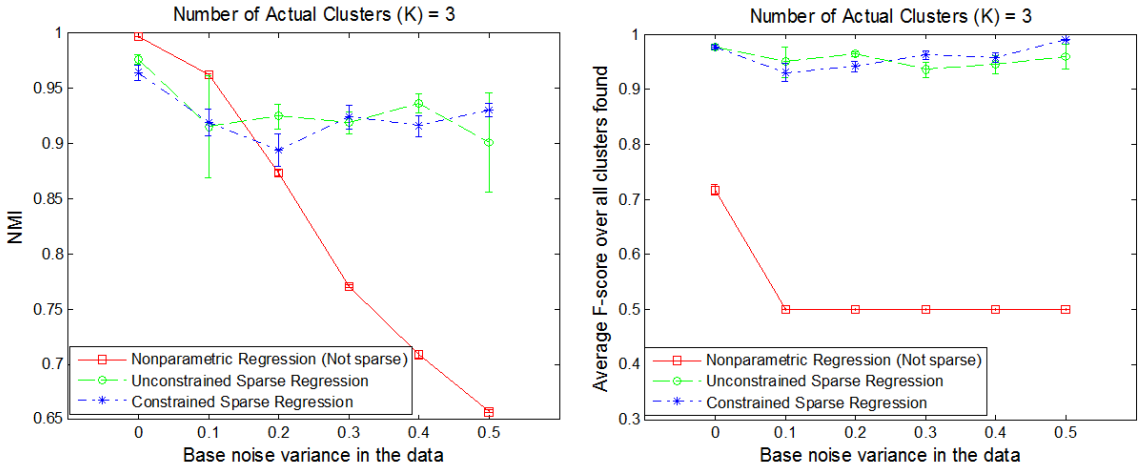

Figure 4. Left panel: ability of nonparametric unregularized and sparse regressions (unconstrained and constrained) to correctly identify clusters (indicated by NMI) with increasing noise. Right panel: ability of nonparametric unregularized and sparse regressions (unconstrained and constrained) to correctly retrieve the sparse structure within each cluster (indicated by average F score).

our method is relatively robust to added noise, a major challenge with the real data sets, especially in terms of correctly identifying the sparse structure.

4.2 Feature selection for downscaling rainfall

A grand challenge in climate science relevant for adaptation and policy remains our inability to provide credible stakeholder-relevant "statistical downscaling", or to develop statistical techniques for more accurate, precise and interpretable high-resolution projections with lower-resolution climate model data (Benestad et al., 2008). Regression models of statistical downscaling (Benestad et al., 2008; Ghosh, 2010) work by first selecting a set of climate variables that have information about the target variable, and then fitting a regression model to predict the target variable at higher resolution. In this application, selecting the right set of predictors is as important as building a prediction model since even a good prediction with a model that is physically not interpretable is less desirable as it may not generalize well. We focus on the feature selection problem for statistical downscaling of annual average rainfall. The use of annual averages reduces the amount of noise in the observed rainfall data, which enables us to examine the robustness of our methods with less ambiguity.

Existence of multiple states or patterns is acknowledged in regression-based statistical downscaling literature for rainfall (e.g., Kannan and Ghosh, 2010) where parametric methods such as k-means were used to find distinct clusters. Here we used our model to simultaneously find clusters, if any, and select features for the purpose of statistical downscaling of station-observed annual average rainfall over two climatologically homogeneous regions over the continental US. Figure 5 shows the climatologically homogeneous regions over the US.

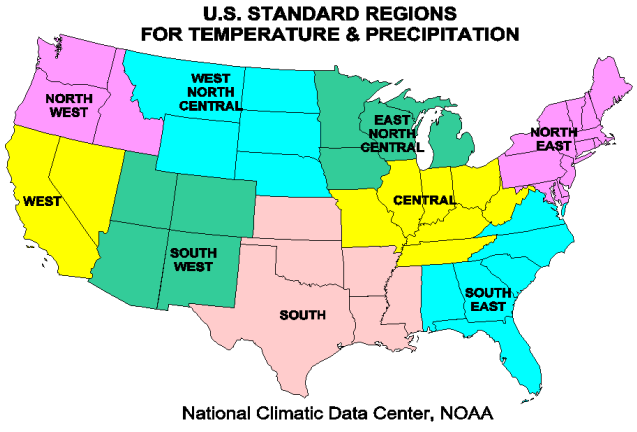

Figure 5. Map showing climatologically homogeneous regions over continental US.

Since rainfall follows a log-normal distribution (Kedem and Chiu, 1987), the target variable we used is logarithm of annual average rainfall. In Fig. 6, we show the distribution of average rainfall over all sites in western US before and after taking the logarithm.

Potential features used can fall in one of two broad categories – local atmospheric variables and large-scale climate indices. Local covariates originate from each station and exhibit both spatial and temporal variability. Annual and seasonal averages of maximum temperature fall in this category along with sea level pressure (SLP), and convective available potential energy (CAPE). A dependence on any of these variables roughly indicates dominance of local convective rainfall in the region. Daily rainfall station data were obtained from US Historical Climatology Network (USHCN) (Easterling et al., 1996). All other features are described in Table 1.

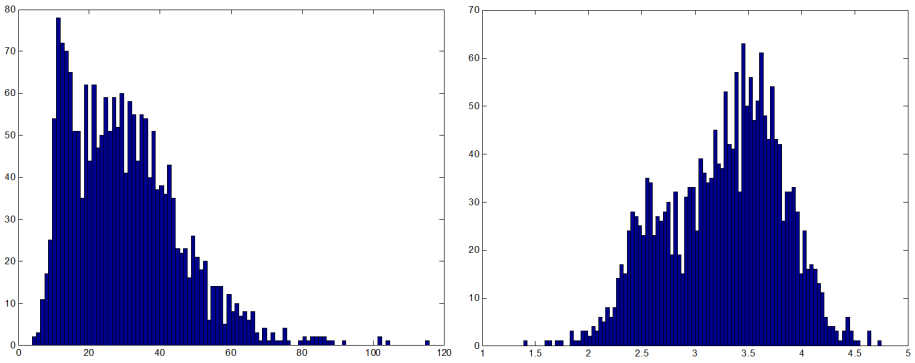

Figure 6. Left panel: distribution of average rainfall over all sites in the western US. Right panel: distribution of average rainfall after transformation.

Table 1. Potential features used for statistical downscaling of rainfall.

Atmospheric (Easterling et al., 1996; Mesinger et al., 2006)	
MATmax	Mean Annual Maximum Temperature
DJFTmax	Mean Winter Maximum Temperature
MAMTmax	Mean Spring Maximum Temperature
JJATmax	Mean Summer Maximum Temperature
SONTmax	Mean Autumn Maximum Temperature (Easterling et al., 1996)
SLP	Sea Level Pressure
CAPE	Convective Available Potential Energy (Mesinger et al., 2006)
Climate indices (NOAA, 2014)	
NAO	North Atlantic Oscillation
EA	East Atlantic Pattern
WP	West Pacific Pattern
EPNP	East Pacific/North Pacific Pattern
PNA	Pacific/North American Pattern
EAWR	East Atlantic/West Russia Pattern
SCA	Scandinavia Pattern
TNH	Tropical/Northern Hemisphere Pattern
POL	Polar/Eurasia Pattern
PT	Pacific Transition Pattern
PDO	Nino 1+2, Nino 3, Nino 3.4, Nino 4
SOI	Southern Oscillation Index
PDO	Pacific Decadal Oscillation
NP	Northern Pacific Oscillation
TNA	Tropical/Northern Atlantic Index
TSA	Tropical/Southern Atlantic Index
WHWP	Western Hemisphere Warm Pool
GlobalMeanTemp	Global Mean Temperature Anomaly (NOAA, 2014)

Climate indices are global variables that represent large-scale signals in climate variables. A list of covariates used for each category is given in Table 1. A dependence on any of these variables roughly indicates rainfall due to large-scale circulation. In addition to these covariates, we have used elevation as a potential feature which falls under none of the above categories. This is the only feature that represents the geography of the region.

We could use the covariates between 1979 and 2011 as SLP and CAPE is available only for that period. Also, if more than 50 % of the daily observations in a year are found to be missing for any covariate at a specific location, we simply discarded all covariates for that year and for that specific location. We averaged monthly climate indices and daily local variables over a year. Finally the annual/seasonal average time-series of predictors for each station were merged for a homogeneous region under consideration. West (CA,

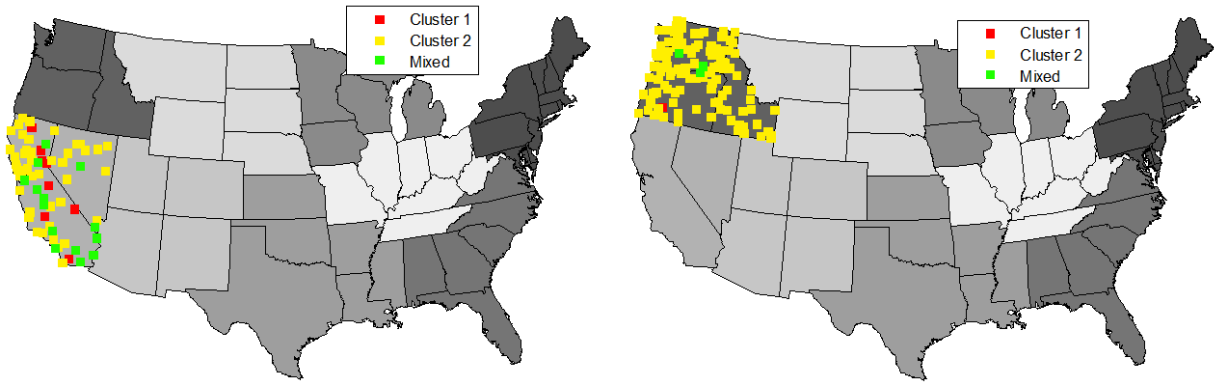

Figure 7. Left panel: location of stations and their cluster membership in the western region. Right panel: location of stations and their cluster membership in the northwestern region.

NV) and northwest (WA, OR, ID) regions are shown by gray shaded areas over the US map in Fig. 7 (left and right panels, respectively).

Results and discussion

We applied spatial "must-link" constraints among pairs of data-points from the same location. Ideally, if there are n points in a cluster, we will be required to put $\binom{n}{2}$ constraints to cover all pairs of data-points. To reduce complexity, initially we kept only those constraints that connect data-points from consecutive years. However, this reduced set of constraints proved to be too restrictive and all data-points tended to merge into a single cluster. So, we kept removing the constraints in an intuitive manner until more than one cluster emerged for a region. We found more than one cluster for all regions except the southern region. We stopped removing constraints until new clusters stopped emerging for a region. Here we show only the clusters in the western and northwestern regions, since the majority of stations were mostly split into obtained clusters in these regions. In other regions, almost all stations had mixed membership. We assign a station to a cluster if more than 80 % of its data-points belong to that cluster.

A quick look at the histogram of target variable (right panel in Fig. 6) also supports the possibility of two distinct rainfall modes in the region. As mentioned earlier, we obtained one sparse linear model for each of the discovered components within a region. Since a non-zero coefficient in the sparse model implies dependence on the corresponding covariate, we can obtain interesting insights about the dependence of average rainfall on various atmospheric and climate indices from the coefficients of the individual sparse models within each cluster. Interestingly, in the northwest region there is only a single member station in the first component that exhibits dependence on the local temperature variables and SLP, whereas the larger cluster shows dependence on a larger number of climate indices. In the western region, the first cluster shows dependence on local temperature variables and the second cluster shows more dependence on large-scale variables. Both clusters show dependence on elevation. While dependence on large-scale indices is not surprising for both these coastal regions due to the known effect of westerlies, dependence of smaller clusters (especially in the northwest) on local variables may hint at the existence of some regional small-scale atmospheric mechanisms. While spatially coherent clusters are more likely to occur in nature, geographical features such as mountains and lakes and even man-made structures such as large dams and reservoirs may abruptly disturb the spatial smoothness of clusters, since their presence may alter the climate pattern of the nearby areas with respect to the surrounding regions. However, before we can build statistical downscaling models, more rigorous statistical and physical analysis is required based on these preliminary insights obtained using our method. The clusters discovered here, and the corresponding covariates, can be utilized to develop individual non-linear prediction models per cluster.

DPMs automatically find the number of clusters K and adapt to varying values of K. However, DPMs prevent the model from "learning" an unnecessarily large value of K if a smaller K is sufficient to describe the model, thus managing complexity. Based on the results of experiments on the synthetic data set shown in Fig. 2, we found that the performance of the method degrades as the number of components K grows larger. We believe it is reasonable to expect that there will only be a limited number of distinct relationships between average rainfall and their covariates when we apply our method at the regional scale. However, even in situations where a large number of relationships exist within a particular region, our method may not be able to identify all of the distinct methods, but it can nevertheless be expected to outperform the use of a single model. The single model will attempt to learn a relationship that is the average of all distinct relations, while our approach will still attempt

to distinguish among major categories of relationships even though some of them may be lumped together.

5 Conclusions

In this paper, we propose a nonparametric Bayesian mixture of sparse regression models for simultaneous clustering and discovery of covariates within each cluster using a DP mixture model. Moreover, our model can accommodate prior knowledge about "must link" constraints between the pair of data-points using a Markov Random Field prior on the cluster membership variables. Our major contribution is to develop an efficient and scalable variational inference algorithm for inference on the fully Bayesian model. We applied our method to both synthetic and real climate data and successfully discovered multiple underlying behaviors in the data. Preliminary results of applying our method to feature selection for statistical downscaling of rainfall show promise towards finding new climate insights with appropriate caveats. Going forward, we would like to incorporate priors for diversity among the clusters in order to discourage merging of close but dissimilar clusters. We intend to extend our model for predictive analysis and build a full-scale statistical downscaling method using the features selected by the current model.

Acknowledgements. This work was funded by the NSF Expeditions in Computing grant "Understanding Climate Change: A Data Driven Approach", award number 1029166. We thank the anonymous referees for their valuable suggestions and comments.

Edited by: V. Kumar
Reviewed by: three anonymous referees

References

Antoniak, C.: Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems, Ann. Stat., 2, 1152–1174, 1974.

Bader, D. C., Covey, C., Gutkowski Jr., W. J., Held, I. M., Kunkel, K. E., Miller, R. L., Tokmakian, R. T., and Zhang, M. H.: Climate Models: An Assessment of Strengths and Limitations, US Climate Change Science Program Synthesis and Assessment Product 3.1, Department of Energy, Office of Biological and Environmental Research, 124 pp., available at: http://pubs.giss.nasa.gov/docs/2008/2008_Bader_etal_1.pdf (last access: 20 July 2014), 2008.

Basu, S., Bilenko, M., Banerjee, A., and Mooney, R.: Probabilistic semi-supervised clustering with constraints, J. Mach. Learn. Res., 71–98, 2006.

Benestad, R., Hanssen-Bauer, I., and Chen, D.: Empirical-Statistical Downscaling, World Scientific Publishing Company, New Jersey, London, 2008.

Bishop, C. and Svenskn, M.: Bayesian hierarchical mixtures of experts, in: Uncertainty in Artificial Intelligence, Morgan Kaufman, San Francisco, CA, 57–64, 2002.

Blei, D. M. and Jordan, M. I.: Variational inference for Dirichlet process mixtures, Bayesian Anal., 1, 121–143, 2006.

Easterling, D. R., Karl, T. R., Mason, E. H., Hughes, P. Y., and Bowman, D. P.: United States Historical Climatology Network (USHCN) Monthly Temperature and Precipitation Data, Tech. rep., Oak Ridge National Laboratory, US Department of Energy, Oak Ridge, Tennessee, available at: http://cdiac.ornl.gov/epubs/ndp/ushcn/ushcn.html (last access: 2 April 2014), 1996.

Ebtehaj, A. M., Foufoula-Georgiou, E., and Lerman, G.: Sparse regularization for precipitation downscaling, J. Geophys. Res., 117, 1–12, doi:10.1029/2011JD017057, 2012.

Ferguson, T.: A Bayesian analysis of some nonparametric problems, Ann. Stat., 1, 209–230, 1973.

Geman, S. and Geman, D.: Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, IEEE T. Pattern Anal., PAMI-6, 721–741, 1984.

Ghosh, S.: SVM-PGSL coupled approach for statistical downscaling to predict rainfall from GCM output, J. Geophys. Res., 115, D22102, doi:10.1029/2009JD013548, 2010.

Greene, A. M., Robertson, A. W., Smyth, P., and Triglia, S.: Downscaling projections of Indian monsoon rainfall using a nonhomogeneous hidden Markov model, Q. J. Roy. Meteorol. Soc., 137, 347–359, 2011.

Hespanha, J. P.: An efficient MATLAB Algorithm for Graph Partitioning, Tech. rep., University of California, Santa Barbara, available at: http://www.ece.ucsb.edu/~hespanha/techrep.html (last access: 2 April 2014), 2004.

Ishwaran, H. and James, L. F.: Gibbs sampling methods for stick-breaking priors, J. Am. Stat. Assoc., 96, 161–173, 2001.

Kannan, S. and Ghosh, S.: Prediction of daily rainfall state in a river basin using statistical downscaling from GCM output, Stoch. Env. Res. Risk. A., 25, 457–474, doi:10.1007/s00477-010-0415-y, 2010.

Kedem, B. and Chiu, L. S.: On the lognormality of rain rate, P. Natl. Acad. Sci. USA, 84, 901–905, 1987.

Knutti, R. and Sedláček, J.: Robustness and uncertainties in the new CMIP5 climate model projections, Nat. Clim. Change, 3, 369–373, doi:10.1038/nclimate1716, 2013.

Kumar, D., Kodra, E., and Ganguly, A. R.: Regional and seasonal intercomparison of CMIP3 and CMIP5 climate model ensembles for temperature and precipitation, Clim. Dynam., 43, 2491–2518, doi:10.1007/s00382-014-2070-3, 2014.

Mesinger, F., DiMego, G., Kalnay, E., Mitchell, K., Shafran, P. C., Ebisuzaki, W., Jović, D., Woollen, J., Rogers, E., Berbery, E. H., Ek, M. B., Fan, Y., Grumbine, R., Higgins, W., Li, H., Lin, Y., Manikin, G., Parrish, D., and Shi, W.: North American regional reanalysis, B. Am. Meteorol. Soc., 87, 343–360, doi:10.1175/BAMS-87-3-343, 2006.

NOAA: Climate Indices: Monthly Atmospheric and Ocean Time Series, available at: http://www.esrl.noaa.gov/psd/data/climateindices/list/, last access: 2 April 2014.

Park, T. and Casella, G.: The Bayesian LASSO, J. Am. Stat. Assoc., 103, 681–686, doi:10.1198/016214508000000337, 2008.

Phatak, A., Bates, B., and Charles, S.: Statistical downscaling of rainfall data using sparse variable selection methods, Environ. Modell. Softw., 26, 1363–1371, doi:10.1016/j.envsoft.2011.05.007, 2011.

Ross, J. and Dy, J.: Nonparametric Mixture of Gaussian Processes with Constraints, The 30th International Conference of Machine Learning, Atlanta, GA, 2013.

Sethuraman, J.: A constructive definition of Dirichlet priors, Stat. Sinica, 4, 639–650, 1994.

Tarjan, R.: Depth-first search and linear graph algorithms, SIAM J. Comput., 1, 146–160, 1972.

Tibshirani, R.: Regression shrinkage and selection via the LASSO, J. Roy. Stat. Soc. B, 58, 267–288, 1994.

Wilby, R., Charles, S., Zorita, E., and Timbal, B.: Guidelines for use of climate scenarios developed from statistical downscaling methods, Tech. Rep. August, Inter-Governmental Panel for Climate Change, available at: http://www.narccap.ucar.edu/doc/tgica-guidance-2004.pdf (last access: 2 April 2014), 2004.